DOCUMENT RESUME

ED 193 293 TM 800 608

AUTHOR Gillmore, Gerald M.
TITLE Student Instructional Ratings: To What Universe Can We Dependably Generalize Results?
PUB DATE Apr 80
NOTE 24p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Aptitude Treatment Interaction; Course Evaluation; *Generalization; Higher Education; *Measurement Techniques; Research Design; *Research Problems; Sampling; *Student Evaluation of Teacher Performance; *Test Validity
IDENTIFIERS *Generalizability Theory

ABSTRACT

The generalizability issue is reviewed through previously published results and by presenting new data. When classes (teacher-course combinations) are used as the object of measurement, studies show very consistent results. However, when either the teacher or the course is treated as the object of measurement, with the other being a facet over which generalization is desired, results are highly inconsistent, especially with regard to the effects of discipline and item content. The hypothesis that discrepancies are largely due to sampling fluctuations in variance component estimates is presented. For teachers and courses as objects of measurement, a closer look at item content and discipline is needed; also necessary, however, is an investigation of course, teacher, and student characteristics, and teaching/learning styles. The pooling of institutional data is one possible way of moving closer to definitiveness. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193293

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. Gillmore
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Student Instructional Ratings: To What Universe Can We Dependably Generalize Results?

Gerald M. Gillmore

University of Washington

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The title of this paper asks, "To what universe can we dependably generalize the results of student instructional ratings?" Archibald MacLeich has stated, accurately I think, that "It is a common human practice to answer questions without truly asking them..." (Riders on the Earth, p. 57). Bowing to his insight, I will risk boring some of you by trying to clarify the question before proceeding to discuss what we know of the answer.

The language of the question correctly implies that I am using generalizability theory (Cronbach, et al., 1972) as a framework, and I think it is uniquely able to productively guide our thinking here. As Cronbach (1976) has stated, "G[generalizability] theory enables you to ask your question better..." (p. 199). (I might parenthetically add that it also allows you to see when the answer obtained is not appropriate to the question asked.) Our question is motivated basically by the fact that in the process of having students rate the quality of their classes, we obtain the responses of specific students, responding to specific items, in regard to the teaching of a specific instructor, who taught a specific course in a specific room at a specific time, etc. If we were happy with these results per se and had no desire to generalize beyond all of those specifics, we would be home free. But, of course, as we use the data to help us make decisions, we become more interested in

TM 800608

Invited address, SIG/Instructional Evaluation, American Educational Research Association Convention, Boston, 1980.

generalizations from the specific to the more general. Typically, we are less interested in the responses of the particular students who rated the course than what the responses of all students who might have taken it would be. Similarly, we are less interested in those particular items than in the item domain from which they were chosen. We may not even be interested in that particular course, but in all the courses the teacher rated may teach. The results we obtained are at best representative of the results we want but cannot get.

Generalizability theory asks how dependably the data we have generalizes to the more ideal set of measurement conditions we would want to base our decisions upon. The "universe of generalization" is basically the set of conditions to which we wish to generalize our obtained measures, e.g., the entire domain from which our items can be considered a random sample or the set of students who might have enrolled in the class.

Generalizability theory allows us to use multi-factor designs in which to collect our data--not being limited to the two of classical test theory. For example, the design under which most of us collect student ratings data is students (s) nested within classes (c) crossed with items (i) or (s:c) x i (see Brennan, 1976). If one wishes to use the resulting data to make decisions, the first question becomes, which of these factors is the object of measurement? The object of measurement simply is that factor about which decisions are to be made. For student ratings, it may be the class, the teacher, the course, or perhaps the department. In a recent Research in Higher Education article Stumpf (1979) used student

ratings to make judgements about the quality of instruction in departments. (Interestingly, all the reliability coefficients cited in support of this usage had as their object of measurement the class--thus rendering them inappropriate in that context).

Once the object of measurement is defined, the remaining factors become facets--or are disregarded as irrelevant. Depending upon the decision to be made, the facets can be considered as random, fixed, or finite. To reduce complexity the finite case will not be discussed here (see Brennan, 1976). Considering a facet as fixed means that the conditions of study are all that are of interest and there is no intention to generalize results to a larger set of conditions for that facet. Considering a facet as random implies that the conditions of the study are a sample from an essentially infinite set, and generalization is to this larger set. For example if our decision leads us to be interested only in the items on the student rating form we would treat the item facet as fixed, but if our decision suggests that we ought to be interested in the entire item domain from which those items are a sample, then the item facet should be considered random.

Oversimplifying somewhat, asking how dependably we can generalize is equivalent to asking how well our obtained measures represent desired measures, and what is desired is dictated by decisions which the measurement will support. Hopefully, this will become much clearer as we subsequently discuss specific studies relating to the question. We can index the dependability of our generalizations by use of the generalizability coefficient. The generalizability coefficient is defined as the ratio of universe score variance (replicable variance) to total observed score

variance. How these quantities are determined depends upon how the universe of generalization is defined. An intuitively satisfying way for me to think of generalizability coefficients is as follows: Consider, theoretically, a complete replication of a study, where you use the same objects of measurement, the same levels of all fixed facets, and a new same-sized random sample of all random facets. The generalizability coefficient is equivalent to the correlation between the replications, across the objects of measurement.

Let us now turn our attention to four specific cases and discuss what can be learned from each of these.

Case 1

Design: (s : c) x i where s = students,
c = classes, and i = items

Object of measurement: Classes

Facets: Students and Items

Example decision: Students selecting classes

Of the cases to be presented, it is this one of which we can speak with most confidence. A number of student ratings studies have been done using this design and a generalizability framework, with surprisingly consistent results (see Kane, Gillmore and Crooks, 1976, and Gillmore, Kane and Naccarato, 1977, for example). In addition, a large number of studies using a classical reliability framework have yielded results consistent with those above. A fairly typical set of variance component estimates, taken from Gillmore, et al., 1977, are presented in Table 1.

Table 1 about here

For case 1, one can view either the student or the item facet as

fixed. Frankly, I know of no interesting application for which the student facet would be appropriately considered fixed. By treating students as fixed, the decision-maker is restricting interest to how the class was rated by the particular students in the class. Yet even now boasts of high inter-item consistency, which is the equivalent reliability coefficient, are found in the student ratings literature. Importantly, if the decision-making context is such that students should be treated as random, but are in fact treated as fixed, grossly inflated reliability or generalizability coefficients can result.

The classical test theory counterpart of treating items as fixed is inter-rater reliability, and I think situations can be described, straining a bit, which suggest treating items as fixed is appropriate. But, almost all serious decisions made about classes would, I think, imply treating both items and students as random facets.

In regard to generalizability coefficients, the variance component associated with the main effect of the object of measurement is always all or part of the universe score variance. Variance components associated with interaction and residual effects are divided by the number of levels of the facet(s) contained in the effect which are to be sampled in making the decision. Hence, generalizability coefficients change as a function of the number of levels to be sampled. In Table 2, generalizability coefficients are presented for differing levels of items and students, considering both facets as random, using the variance component estimates of Table 1.

Table 2 about here

Briefly, one can see that the number of students rating the class

has a greater impact on dependability than the number of items. Increases in the number of items used beyond five have little effect on dependability. The dependability no matter how many items is used tends to be low with 10 or fewer students.

Case 2

Design: (s : o) x c x i where o = occasions

Object of Measurement: Classes

Facets: Occasions, students, and items

Example decision: Quality of small classes

Case two can be introduced with the question, what can one do to get a dependable result if a class is small? Common sense would suggest that one should sum the results of several offerings of the course by the same instructor. Smith (1979a) presented data relevant to the problem. The data are from twelve Educational Psychology teachers, each teaching a same course for three successive semesters. Smith chose to treat the occasions factor as crossed within his design: he could have treated occasions as nested within classes with equivalent results.

The variance component estimates of the study are found in Table 3.

Table 3 about here

The variance components associated with the occasions main effect, and all interactions containing occasions are estimated to be zero in magnitude. Hence, according to Smith's results, varying levels of the occasions facet has no direct effect on resulting dependability estimates. But it has the indirect effect of increasing the number of students. Since students are nested within occasions, any increase in the number of occasions automatically is accompanied by an increase in the number of

students and hence an increase in the dependability of the results of combining occasions or sections. Common sense prevails!

Case 3

Design: (s : c : t) x i where c = courses and
t = teachers

Object of measurement: Teachers

Facets: Courses, students, and items

Example decision: Tenure

Under case 3, the focus changes from decisions about classes to decisions about teachers. The results discussed above would apply to decisions about teachers only if the specific courses taught were of no moment. Most of us would not buy that assumption without evidence. Hence, the need to separate the teacher effect and the course effect from the class effect. A design in which courses are crossed with teachers would constitute an ideal means to view this question; however, within higher education seldom do the same set of instructors teach the same set of courses. A less direct means to untangle teacher and course effects first used by Gillmore, Kane and Naccarato (1978) and later Smith (1979b) is draw samples of teachers whose teaching of two different courses was rated, and rated courses which were taught by two different teachers. Hence, under the former design, courses are nested within teachers. (Students are nested within courses and crossed with items).

The Gillmore et al, (1978) and the Smith (1979b) studies differed in some potentially important respects. Gillmore et al, used 42 teachers, each teaching two courses, with 13 students per class, and 11 undifferentiated items. The 42 classes were sampled from throughout the University of Washington, with 14 from the Humanities, 14 from the Social Sciences,

and 14 from the Physical Sciences. Smith used 20 teachers, each teaching two courses, all from the Educational Psychology Department at the University of Wisconsim, Milwaukee. He had 10 items which were instructor oriented, 8 items which were course oriented, and 10 students per class.

In regard to item content, Smith correctly pointed out that decisions to be made about instructors should be made using items which are instructor related. In Gillmore et al, the item domain was more loosely defined.

The variance component estimates for Gillmore et al, and for Smith are found in Table 4. I should like to focus on important differences in

Table 4 about here

the magnitudes of the teacher and the course within teacher effects. Because of the nested design, the course effect and the teacher by course interaction are confounded in the course within teacher (c : t) effect. For Gillmore et al, the c : t effect has somewhat more variance attributed to it than the teacher effect. But, for Smith's instructor items, course items, and all items, the teacher effect is much larger than the course within teacher effect. In fact, in all three cases, but particularly for the instructor items, Smith's variance component estimate for c : t is relatively negligible. This result implies little or no variance attributable to either the course or the course by teacher interaction.

The Gillmore et al, Smith differences can be highlighted with regard to resulting generalizability coefficients. First, consider the case in

Table 5 about here

which one wants to generalize over courses, i.e., treat courses as random. Gillmore et al, would suggest that for classes of approximately 20 students, ratings should be done on 5 to 10 courses before a dependable average for

a teacher emerges. Smaller classes need closer to 10 courses. Smith, in contrast, would require only one or two classes for either his instructor items or all items, with moderately sized classes. Small classes might require more. In other words, in Smith's results courses as a facet show results approximately equivalent to occasions as discussed earlier.

If courses are treated as fixed, the results of the two studies in regard to generalizability coefficients are much more similar, simply because the course within teacher variance becomes part of the universe score variance rather than error variance.

Before discussing these important discrepancies, I would like to turn our attention to the fourth and final case.

Case 4

Design: (s : t : c) x i

Object of measurement: Courses

Facets: Teachers, students, and items

Example decision: Curricular revision

In design, case four is the flip side of case three. It speaks to those situations in which decisions are to be made about courses, e.g., History 100. Perhaps a curriculum committee might wish to make use of student opinion in deciding the fate of a given course.

Both Gillmore et al. and Smith included this design in the studies mentioned earlier. However, the discrepancy between the two studies is even greater here. Table 6 lists the variance component estimates found by Gillmore, and by Smith for his instructor items, course items, and all items. Again let us focus our attention on the course effect and the teacher within course effect. The variance component estimates of Gillmore et al, are approximately equivalent to those of Smith for the instructor items.

Table 6 about here

But, there is a dramatic difference in regard to Smith's course items. Whereas Gillmore et al, were led to conclude that one could not obtain dependable data for courses, when generalizing over teachers, for any reasonable sized number of teachers, Smith can conclude that by using course related items, dependable data emerges when two to five different teachers teach the course, depending upon the class sizes. (see Table 7)

Table 7 about here

Inconsistencies

There seem to be at least four possible explanations for the discrepancies. First, the studies were done at two institutions which may differ in some important respects. I am inclined to place less credence on this explanation since I have seen no evidence that differences among institutions lead to the discrepancies noted above.

More promising in explaining differences is the fact that Gillmore et al, effectively sampled the entire university, while Smith restricted attention to the Educational Psychology Department. In regard to the dependability of generalization, we know little of the importance of discipline differences.

Of course, Smith's differentiation of items also suggests a third possible point of difference. Intuitively, one would suspect that items which are course related would yield greater differentiation among courses. Yet taking the high intercorrelations among items across classes which are consistently found for student ratings, it is hard to believe that the magnitude of the differences found is attributable to item content

alone. Furthermore, comparisons of Smith's results for all items and Gillmore's results still show important differences.

I would like to introduce my fourth possible reason by discussing an aspect of Smith's results which trouble me. In Table 8, I have placed the variance component estimates of the course items for Smith's c : t and t : c studies side by side.

Table 8 about here

Note that the variance component estimates are very similar for the five lower components, as should be the case. However, the variance component for c : t is estimated to be quite small by the data of the design with teacher as the object of measurement. Since

$$\hat{\sigma}^2 (c : t) = \hat{\sigma}^2 (c) + \hat{\sigma}^2 (ct)$$

this result implies that the variance attributable to both of the sources is small. Now look at the results of the design in which teachers are nested within courses. The variance component for courses is relatively large, but the confounded estimate for teachers and the teacher by course interaction is small. Thus, the first data set indicates little variance for the course effect, the second indicates much more. Likewise the second data set indicates little variance for the teacher effect, the first much more. I submit the two results are directly inconsistent. They cannot both be correct--even if we assume zero teacher by course interaction.

To what can we attribute this inconsistency? I would <u>hypothesize</u> that it is a result of sampling fluctuations in regard to variance component estimates. Smith (1978) in presenting his important research on the subject correctly stated that "...little is known about the sampling characteristics of estimates for variance components" (p. 322).

Formulae for determination of the variance of variance component estimates are available. Unfortunately these formulae are in terms of population values for variance components, along with degrees of freedom, and, hence, of somewhat limited value. In addition, formulae for higher order effects in complex designs are themselves complex and lengthy. For example, for the (s : c : t) x i design, I worked out the formula for the variance of the variance component estimate of the teacher main effect to contain on the order of eighteen terms--depending upon how one chooses to count. Producing this gawd-awful formula here would serve little purpose, but let me briefly indicate one important aspect of it.

The formula for estimating the variance component for the teacher main effect is as follows, where MS(T) is the mean square for the teacher effect:

$$n_c \cdot n_s \cdot n_i \cdot \hat{\sigma}^2(T) = MS(T) - \left[n_s \cdot n_i \cdot \hat{\sigma}^2(c : t) + n_c \cdot n_s \cdot \hat{\sigma}^2(TI) + n_s \cdot \hat{\sigma}^2(CI : T) + n_i \cdot \hat{\sigma}^2(S:C:T) + \hat{\sigma}^2(SI:C:T) \right]$$

Using Smith's sample sizes for illustrative purposes, each of the components of the equations is estimated with the degrees of freedom found in Table 9.

Table 9 about here

Note that the degrees of freedom for estimating the mean square for teachers is relatively small. But it is further compromised by sampling fluctations in variance components estimates lower in the hierarchy, especially the variance component estimate for courses nested within teachers. This latter effect has relatively few degrees of freedom as well.

Additional Data

To reiterate, I suggested discrepancies between the results of Smith

and Gillmore et al. might be attributable to disciplinary differences, item content, or sampling fluctuations. To attempt to shed light on the question, I drew data from three departments at the University of Washington: Electrical Engineering, English, and Psychology. These were departments making the greatest use of our student ratings system as well as being nicely divergent. Using the course nested within teachers and teachers within courses designs, I was, of course, restricted to teachers who had been rated within two courses, and rated courses taught by two different teachers. Even with drawing data from an entire academic year, excluding Summer, my sample sizes with regard to courses and teachers were rather small.

The resulting variance component estimates for the teacher effect and the course-within-teacher effect, and the course effect and the teacher within course effect are presented in Table 10. The remaining variance component estimates are not presented because they turned out to be very consistent and not relevant to the arguments I intend to present here. One can also note that variance components estimates are presented for all items (n_i = 11), for course related items (n_i = 4) and for instructor related items (n_i = 4). The content of these items is found in Table 11.

Tables 10 and 11 about here

As I stated earlier, the purpose of this study was to help clarify the importance of the discipline in terms of the dependability with which certain generalizations can be made, and to help clarify the role of item content. Careful reading of Table 10 clearly indicates that I was thwarted. One has to view the presence of negative variance component estimates for

courses in Electrical Engineering and for teachers in Psychology with alarm. We all know variances cannot be negative. Their estimate can be negative, but only by sampling error. Cronbach et al. (1972) suggested that negative variance components be treated as zero, which is reasonable advice for a "minor" component, However, the teacher is the object of measurement and the teacher effect variance component is all or part of the universe score variance. A zero variance component would indicate no generalizability when the course, student, and item facets are treated as random. Likewise for the course effect in the teacher within courses design. I just do not believe this result. It flies in the face of all past research and a lot of intuition.

The negative variance component estimates, and surely other inconsistencies in Table 10, must be attributable at least in part to sampling errors. Richard Light in a recent Educational Researcher Article (1979) used a wonderful quote from Mark Twain. "The thirteenth stroke of a clock is not only false of itself, but casts grave doubts on the credibility of the preceding twelve." To me, the data I have just presented offered something of a thirteenth stroke, however, I do not feel we have to completely disregard the preceding metaphorical twelve. Rather, as we try to refine our understandings of the issue addressed in the paper we must perhaps be more aware of the problems associated with arriving at definitive answers.

I continue to view as important questions surrounding the dependability with which we can make certain generalizations when using student ratings results for decision-making. And we will have to ask similar questions of other indicators of teaching quality as their use becomes more pervasive. There is a certain irony in my conclusion that some of our results are suspect because of too little data--especially with my access to the data

of one of the largest single campus programs in the country.

In fact, I believe adequate data are out there to get a handle on the measurement issues presented here; particularly issues dealing with decisions concerning teachers, courses, and the interaction between the two. Our understanding of classes as the object of measurement is relatively good--unfortunately, close analysis reveals that this object of measurement is not very interesting. Few important decisions are made about classes, that is, specific course-teacher combinations.

For teachers and courses as objects of measurement, we need to take a closer, more thorough look at item content and discipline, of course. We must also look at other course characteristics, such as course level, teacher characteristics, teaching style, and even student characteristics, such as learning style. To further our understandings we are going to have to be creative and cooperative. I could not find adequate numbers of pairs of teachers and courses within departments from my institution. Perhaps if several institutions could pool data, we could move closer to definitiveness. Also, we are going to have to demand replications from ourselves and others.

My conclusions are considerably weaker than I had anticipated when originally agreeing to present this paper. Perhaps my goals have become more grandiose, namely, to peak your interest and gain some measure of intelligent commitment to seek true solutions.

References

Brennan, R. L. Generalizability analyses: principles and procedures. ACT Technical Bulletin, No. 26, Iowa City: The American College Testing Program, 1977.

Cronbach, L. J. On the design of educational measures. In D. N. M. de Gruijter and L. J. T. van der Kamp (eds.), Advances in Psychological and Educational Measurement. New York: Wiley, 1976.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of generalizability for scores and profiles. New York: John Wiley and Sons, 1972.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. The generalizability of student instructional ratings: general theory and application to the University of Washington Instructional Assessment System. EAC Reports, No. 76-14, Seattle: Educational Assessment Center, University of Washington, 1976.

Gillmore, G. M., Kane, M. T. & Naccarato, R. W. The generalizability of student ratings of instruction: estimation of the teacher and course components. Journal of Educational Measurement, 1978, 15, 1-13.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: the generalizability of class means. Journal of Educational Measurement, 1976, 13, 171-184.

Light, R. J. Capitalizing on variation: how conflicting research findings can be helpful for policy. Educational Researcher, 8, October, 1978, 7-11.

MacLeich, A. Riders on the Earth. Boston: Houghton Mifflin, 1978.

Smith, P. L. Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Statistics, 3, 1978, 319-346.

References continued

Smith, P. L. The stability of teacher performance in the same course over time. Research in Higher Education, 11, 1979a, 153-165.

Smith, P. L. The generalizability of student ratings of courses: asking the right questions. Journal of Educational Measurement, 16, 1979b, 77-87.

Strumpf, S. A. Assessing academic program and department effectiveness using student evaluation data. Research in Higher Education, 11, 1979, 353-363.

Table 1

Variance Component Estimates for the (S : C) x 1 Design

(Taken from Gillmore, et al, 1977, Table 3, p. 25.)

Source	Variance Component Estimate
C	.13
S : C	.43
I	.01
CI	.05
SI : C	.46

C = Classes
S = Students
I - Items

Table 2

Estimated Case 1 Generalizability Coefficients with Items and Students as Random Facets (Taken from Gillmore, et al., 1977, Table 4, p. 26)

		Number of Items				
		1	2	5	10	25
	1	.12	.16	.20	.21	.22
Number of	5	.36	.45	.53	.56	.59
Students	10	.48	.59	.68	.71	.74
	25	.60	.72	.81	.84	.87
	100	.69	.80	.90	.93	.95

Table 3

Variance Component Estimates for the (s : o) x c x i Design (Taken from Smith, 1979a, Table 3, p. 162)

Source	Variance Component Estimate
T	.11
O	.00
I	.12
S : O	.00
TO	.00
TI	.08
OI	.00
TS : O	.37
IS : O	.00
TOI	.00
TIS : O	.51

Table 4

Variance Component Estimates for the (s : c : t) x i Design (Taken from Gillmore et al, 1978, Table 2, p. 7 and from Smith, 1979b, Table 1, p. 82)

		Smith		
Source	Gillmore et al	Instructor Items	Course Items	All Items
T	.08	.21	.12	.17
C : T	.12	.03	.04	.03
S : C : T	.49	.33	.35	.31
I	.02	.13	.08	.14
TI	.03	.04	.07	.06
CI : T	.03	.02	.03	.03
SI : C : T	.50	.49	.45	.51

Table 5

Estimated Case 3 Generalizability Coefficients with Items and Students as Random Facets (Taken from Smith, 1979, Table 2, p. 83)

				Smith (1979)					
		Gillmore et al. (1978)		Instructor Items		Course Items		All Items	
	n'_c \ n'_s	5	20	5	20	5	20	5	20
Courses Random	1	.26	.35	.64	.78	.46	.59	.59	.73
	2	.40	.53	.78	.87	.61	.72	.73	.83
	5	.62	.71	.89	.93	.78	.84	.86	.91
	10	.82	.84	.93	.96	.85	.89	.91	.93
Courses Fixed	1	.65	.87	.74	.91	.66	.85	.72	.88
	2	.71	.93	.84	.94	.75	.89	.81	.92
	5	.81	.95	.92	.96	.85	.92	.89	.94
	10	.95	.97	.95	.97	.89	.93	.93	.95

$n'_i = 10$

Table 6

Variance Component Estimates for the (s : t : c) x i Design (Taken from Gillmore et al, 1978, Table 4, p. 9 and from Smith, 1979b, Table 3, p. 84)

		Smith		
Source	Gillmore et al	Instructor Items	Course Items	All Items
C	.01	.02	.14	.08
T : C	.16	.21	.05	.12
S : T : C	.48	.30	.34	.29
I	.01	.12	.07	.15
CI	.01	.00	.07	.02
TI : C	.04	.07	.04	.07
SI : T : C	.47	.45	.45	.48

Table 7

Estimated Case 4 Generalizability Coefficients with Items and Students as Random Facets (Taken from Smith, 1979, Table 4, p. 86)

				Smith (1979)					
		Gillmore et al. (1978)		Instructor Items		Course Items		All Items	
	n'_t \ n'_s	5	20	5	20	5	20	5	20
Teachers Random	1	.04	.05	.05	.06	.51	.64	.29	.35
	2	.07	.10	.10	.12	.66	.77	.45	.52
	5	.16	.20	.23	.26	.81	.87	.66	.72
	10	.27	.33	.37	.42	.88	.91	.79	.83
Teachers Fixed	1	.61	.85	.75	.90	.68	.86	.73	.89
	2	.62	.85	.76	.91	.78	.90	.79	.91
	5	.65	.86	.80	.92	.87	.93	.86	.94
	10	.69	.86	.83	.94	.91	.94	.91	.96

$n'_i = 10$

Table 8

Estimated Variance Components for Smith's Course Items

(S : C : T) x I		(S : T : C) x I	
Source	Estimated Variance Component	Estimated Variance Component	Source
T	.12	.14	C
C : T	.04	.05	T : C
S : C : T	.35	.34	S : T : C
I	.08	.07	I
TI	.07	.07	CI
CI : T	.03	.04	TI : C
SI : C : T	.45	.45	SI : T : C

Table 9

Degrees of Freedom for Estimating Variance Components ($n_t = 20$, $n_c = 2$, $n_s = 10$, $n_i = 8$)

Component	Degrees of Freedom
MS (T)	19
$\hat{\sigma}^2$ (C : T)	20
$\hat{\sigma}^2$ (TI)	133
$\hat{\sigma}^2$ (CI : T)	140
$\hat{\sigma}^2$ (S : C : T)	360
$\hat{\sigma}^2$ (SI : C : T)	2520

Table 10

Selected Variance Component Estimates for Data from Courses within Electrical Engineering, English, and Psychology

Department	Items	Courses within Teachers		Teachers within Courses	
		$\hat{\sigma}^2$(T)	$\hat{\sigma}^2$(C:T)	$\hat{\sigma}^2$(C)	$\hat{\sigma}^2$(T:C)
Electrical Engineering	All	.19	.06	Negative	.19
$n_t = 15$, $n_s = 7$	Instructor	.12	.02	Negative	.30
$n_c = 7$, $n_s = 7$	Course	.24	.09	Negative	.18
English	All	.06	.06	.01	.08
$n_t = 13$, $n_s = 6$	Instructor	.01	.11	.03	.09
$n_c = 9$, $n_s = 11$	Course	.07	.04	Negative	.09
Psychology	All	Negative	.20	.20	.06
$n_t = 13$, $n_s = 6$	Instructor	Negative	.29	.16	.13
$n_c = 13$, $n_s = 7$	Course	Negative	.06	.28	.03

Table 11

Item Content

Instructor related items

The instructor's contribution to the course was:

The instructor's effectiveness in teaching the subject matter was:

Instructor's interest in whether students learned was:

Clarity of student responsibilities and requirements was:

Course related items

The course as a whole was:

The course content was:

Relevance and usefulness of course content is:

Reasonableness of assigned work was:

Remaining items

Use of class time was:

Amount you learned in the course was:

Evaluative and grading techniques (tests, papers, projects, etc.) were: